TEMA 3. INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA

1. Concepto de estadística

1.1. De acuerdo con el fin del análisis: estadística descriptiva vs. inferencial
1.2. De acuerdo con la metodología aplicada: estadística paramétrica vs. no paramétrica
1.3. Según el número de variables que atiende el análisis: estadística univariada, bivariada y multivariada

2. Conceptos básicos

3. Preparación y presentación de los datos

3.1. Conceptos básicos
3.2. Tipos de presentación tabular
3.3. Representación gráfica de los datos

4. Medidas de tendencia central

4.1. Moda (Mo)
4.2. Mediana (Md)
4.3. Media aritmética ($\overline{X}$)
4.4. Comparación entre la media, la mediana y la moda

5. Medidas de variabilidad y dispersión

5.1. Amplitud Total (AT)
5.2. Cuantiles
5.3. Varianza (s^2) y desviación típica (s)
5.4. Coeficiente de variación (CV)

6. La curva normal

6.1. Concepto y características de la curva normal
6.2. Concepto de asimetría. Tipos
6.3. Puntuaciones transformadas

Bibliografía

1. CONCEPTO DE ESTADÍSTICA

La estadística es la disciplina que intenta sacar conclusiones de estudios empíricos mediante la utilización de modelos matemáticos. Sirve como nexo entre los fenómenos reales y los modelos matemáticos.

Existen diversos criterios de clasificación de los métodos y técnicas estadísticas:

1.1. De acuerdo con el fin del análisis: Estadística descriptiva vs inferencial

La *Estadística descriptiva* tiene como fin describir un conjunto de datos. Es decir, recoge, organiza y sintetiza la información.

La *Estadística Inferencial* tiene como fin hacer inferencias a partir de una muestra sobre una población. Es decir, cómo se tiene que realizar el proceso de extrapolación de los resultados.

1.2. De acuerdo con la metodología aplicada: Estadística Paramétrica vs Estadística no Paramétrica

En la *Estadística Paramétrica*, que es la más exigente, las variables son cuantitativas y deben cumplirse tres supuestos paramétricos:

- Los datos obtenidos se ajustan a la distribución normal.
- Homogeneidad de la varianza (medida de dispersión) entre grupos.
- Las medidas son, mínimo, de intervalo (ej. CI).

En la *Estadística no Paramétrica* se utiliza a partir de escalas nominales u ordinales con variables cualitativas, o bien, cuando no se cumple alguno de los tres supuestos anteriores.

Por lo tanto, es muy importante realizar un análisis previo de los datos, de qué clase son, cómo se distribuyen y qué tipo de estadísticos se pueden aplicar.

1.3. Según el número de variables que atiende el análisis: Estadística Univariada, Bivariada y Multivariada.

La *Estadística Univariada* incluye todas las técnicas que hacen referencia a la descripción e inferencia de una sola variable.

La *Estadística Bivariada* incluye todas las técnicas que hacen referencia a la descripción de dos variables

La *Estadística Multivariada* se utiliza al trabajar con tres o más variables. Es la más utilizada en Educación. Su análisis se ha facilitado con la aplicación de programas informáticos como el SPSS.

2. CONCEPTOS BÁSICOS

A) Población: Todo el conjunto de elementos, finito o infinito, que tiene una o varias características que satisfacen el objeto de estudio de una investigación. Ej. Universidad de Alicante.

- *Censo:* Está directamente relacionado con la población. Es un listado de los elementos que componen una población.

Muestra: Es cualquier subconjunto de una población y, para que sea válida, ha de ser representativa de la población porque se va a trabajar con ella y las conclusiones se van a extrapolar a la población. Ej. 300 alumnos de la Universidad de Alicante.

B) Parámetro: Es cualquier función definida a partir de los valores numéricos de una población. Se representan con letras griegas.

μ = media

σ = desviación típica

Estadístico: Es cualquier función calculada sobre los valores numéricos de una muestra (media, moda, mediana, varianza...). Todos ellos permiten describir en forma simplificada al conjunto de datos obtenidos en la muestra.

$\overline{X}$, M = media

S, DT = desviación típica

En definitiva, lo que en la investigación interesa es describir las poblaciones. Pero debido a que suelen ser muy grandes y su conocimiento es costoso, la Estadística Inferencial se encarga de estimar los parámetros a partir de los correspondientes estadísticos.

3. PREPARACIÓN Y PRESENTACIÓN DE DATOS

3.1. Conceptos básicos

Tabular: Es clasificar la información de forma resumida mediante una tabla.

Tabla: Conjunto de clases o modalidades

Clase: Agrupaciones de distintos elementos que siguen un criterio (exhaustivas, excluyentes, definidas).

Frecuencia absoluta (F): número de observaciones que aparece en cada clase o modalidad.

Frecuencia relativa (F_r): es igual al cociente entre las frecuencias absolutas y el número total de datos.

$$F_r = \frac{f}{N}$$

Porcentajes: columnas de las frecuencias relativas multiplicadas por 100. Tiene la misma función que las frecuencias relativas.

$$\% = F_r * 100$$

	F	**F_r**	**%**
Bomberos	30	0’3	30
Médicos	20	0’2	20
Albañiles	40	0’4	40
Psicopedagogos	10	0’1	10
	100	1’00	100%

<u>Frecuencia acumulada</u> (F_a): Indica el número de casos comprendidos en un intervalo o por debajo del mismo. La frecuencia acumulada no se puede conocer en variables cualitativas en escala nominal.

CI	**F**	**F_a**
121 - 130	5	24
111 - 120	7	19
101 - 110	9	12
91 - 100	3	3

3.2. Tipos de representación tabular

1. Distribución categórica

Se utiliza con variables cualitativas en escala nominal. Lo único que puede determinarse con estos datos es la frecuencia de aparición de sus modalidades (n = 110, 40 son de modalidad 1-hombre- y 70 de la modalidad 2 –mujeres-).

ESTUDIOS	**F**	**F_r**	**%**
Alumnos de Psicopedagogía	92	0’47	47%
Alumnos de Derecho	41	0’21	21%
Alumnos de Magisterio	62	0’32	32%

N = 195

Generalmente, la representación gráfica de este tipo de datos se realiza mediante el diagrama de barras y el ciclograma.

2. Distribución por rangos:

Se puede utilizar para variables cuantitativas o cualitativas en escala ordinal. Consiste en ordenar todos los elementos de la muestra y asignarles un rango u orden (categorías profesionales, curso académico, notas de clase).

Para organizar estos datos se utilizan tablas de frecuencias incluyendo las frecuencias, las frecuencias relativas y los porcentajes o frecuencias acumuladas.

Ejemplo 1: En los diez centros de Educación Primaria de una localidad se han diagnosticado 622 niños/as como superdotados. Se pone en prueba un programa de enriquecimiento para esta muestra. Después de la aplicación de dicho tratamiento se pasa una prueba para analizar su eficacia. Los resultados obtenidos nos indican que han conseguido una mejoría máxima 134 niños/as, 212 moderada, 129 leve y en 147 sujetos la mejoría ha sido nula.

MEJORÍA	**F**	F_r	**%**	F_a	F_{ra}	$\%_a$
Máxima (4)	134	0’2154	21’54	622	1’0000	100’00
Moderada (3)	212	0’3408	34’08	488	0’7846	78’46
Leve (2)	129	0’2074	20’74	276	0’4437	44’37
Nula (1)	147	0’2363	23’63	147	0’2363	23’63
	622	1’0000	100’00			

El diagrama de barras es la representación gráfica más común para estos datos. En el eje de abcisas (x) deben ser colocadas las clases de manera ordenada y en el eje de ordenadas (y) se colocan las frecuencias.

3. Distribución con intervalos:

Se utiliza cuando tenemos variables cuantitativas tanto discretas como continuas (escala de intervalo y de razón). Las clases que forman la tabla están definidas numéricamente, de ahí que se les llame intervalos.

a) Conceptos básicos de las distribuciones de intervalos

♦ ***Intervalo:*** Distancia entre dos valores. Cualquier intervalo viene definido por dos valores llamados límites de intervalo. Uno es el límite inferior y el otro el superior.

Existen dos tipos de límites:

- Aparentes: son aquellos que tenemos en una distribución cuando existe discontinuidad entre un intervalo y el siguiente.

- Reales: aquellos que no presentan discontinuidad entre un intervalo y el siguiente. Se calcula:

  * El límite real superior: se toma el límite aparente superior más media unidad de media (+ 0’5).

  * El límite real inferior es igual al límite aparente inferior menos media unidad de medida (- 0’5).

▪ ***Marcas de clase:*** Es el valor que sirve para representar al intervalo. Es decir, es el punto medio del intervalo y se calcula:

$$X_m = \frac{\text{lim. Sup. } + \text{ lim. Inf.}}{2}$$

▪ ***Amplitud del intervalo:*** Es la cantidad de valores que recoge el intervalo (tamaño del intervalo). Hay de dos tipos:

- Unitarios: tienen una amplitud de 1
- No unitarios: su amplitud es superior a 1

X	F
10	2
9	2
8	3
7	4
6	5
5	8

La amplitud en este caso sería 1.

106-110
101-105
96-100
91-95

Cálculo:

A = Lim. Real sup. – lím. Real inferior .... A = 95’5 – 90’5 = 5

A = Lím. Apar. Sup. – lím. Apar. Inf. + 1 .... A = 95 – 91 + 1 = 5

- ***Tipos de intervalos***

1) Intervalos abiertos: carecen de límite superior o de límite inferior
2) Intervalos cerrados: tienen los dos límites

El inconveniente de los intervalos abiertos es que no podemos calcular las marcas de clase puesto que desconocemos uno de los límites. Estos se soluciona con otra clasificación:

1) Intervalos iguales: todos los intervalos presentan la misma amplitud.
2) Intervalos desiguales: cuando los intervalos no tienen la misma amplitud.

111 – 130
86 - 95

Construcción de tablas de intervalos

1. Se calcula la amplitud total incluyente o Rango total incluyente, que es la distancia entre el valor máximo y el valor mínimo.

$$AT_i = \text{Pto. Máx.} - \text{Pto. Mín.} + 1 \text{ unidad de medida}$$

2. Ver el número óptimo de intervalos. La precisión será mayor cuando mayor sea el número de intervalos. Sin embargo, es una decisión del investigador.

- Si N es menor a 100 conviene que el número de intervalos no exceda del valor de la raíz cuadrada de N.

  $\sqrt{85}$ = 9'219 ≈ 9. El número de intervalos es 9

- Si N es mayor que 100 el número de intervalos debe oscilar entre 10 y 20

3. Calcular la amplitud del intervalo:

$$A = \frac{AT_i}{\text{Nº de intervalos}}$$

4. Establecimiento de los límites, especialmente el límite aparente inferior que será el punto mínimo.

Ejemplo:
14,10,13,8,8,7,4,5,14,13,10,11,5,8,9,13,11,11,11,10

1) $AT_i = 14 - 4 + 1 = 11$
2) Nº de intervalos aproximado: Aunque $\sqrt{20}$ = 4'7 tomaremos 6.
3) A = 11/6 = 1'83 ≈ 2

4)

14 - 15	2
12 – 13	3
10 – 11	7
8 – 9	4
6 – 7	1
4 - 5	3
	20

- ***Tablas estadísticas múltiples:*** Son las que se utilizan cuando queremos estudiar la distribución conjunta de dos o más variables. Se trata de variables cruzadas y sirve tanto para variables cualitativas como cuantitativas.

Ejemplo:

V_1 = Profesión

V_2 = Sexo

V_3 = Edad

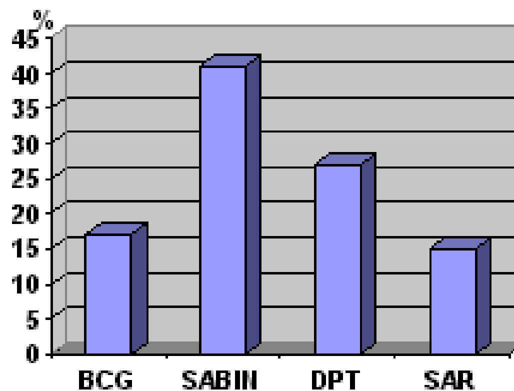
	Menos de 40 años		**Más de 40 años**	
	Hombres	**Mujeres**	**Hombres**	**Mujeres**
Bombero	15	3	7	1
Astronauta	7	1	5	28
Psicopedagogo	27	5	14	3
Otros	4	14	5	9

Para la representación gráfica se puede utilizar el histograma o el polígono de frecuencias.

3.3. Representación gráfica de los datos

Se suele representar siempre mediante coordenadas cartesianas. El eje de abcisas (x) tiene los valores que estamos estudiando y el eje de ordenadas (y) representa las frecuencias.

Diagrama de barras: Es un gráfico formado por un conjunto de barras o rectángulos, que se dibujan sobre unos ejes de coordenadas. Cada una de las barras representa una categoría y deben estar separadas entre ellas por espacios en blanco. La anchura de las barras es elegida arbitrariamente. Normalmente debe guardar una relación 3/5, es decir, que si el eje de ordenadas mide 9 cm., el eje de abcisas debe medir 15 cm.

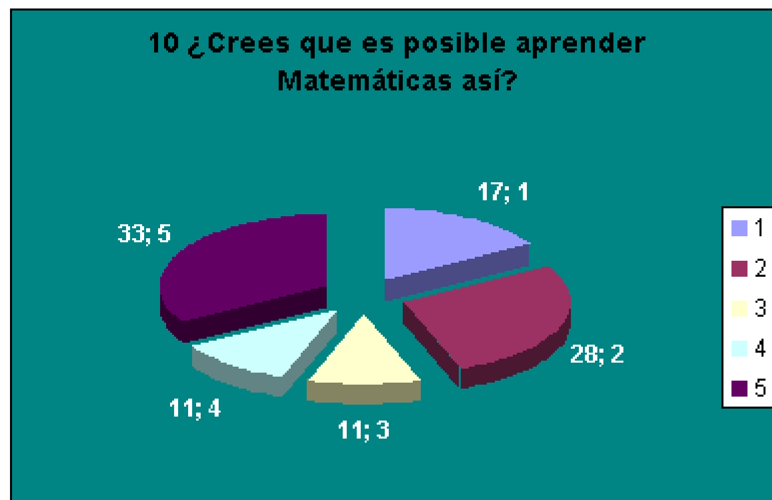


Ciclograma o diagrama de sectores: Está formado por un círculo subdividido en sectores por algunos de sus radios. La superficie o área de tales sectores ha de ser proporcional a la frecuencia (normalmente porcentajes).

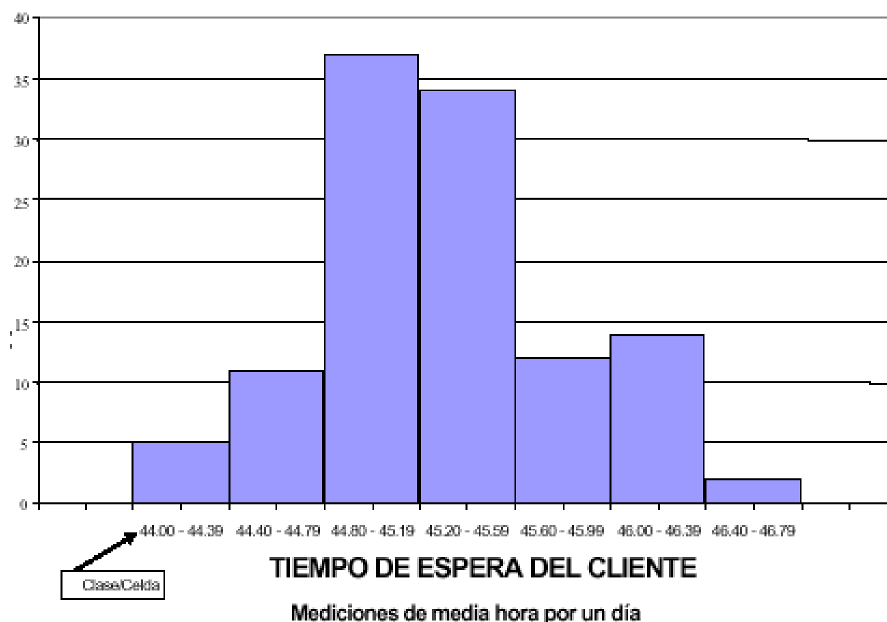
Para el cálculo del número de grados que ocupa cada sector se aplica una regla de tres:

$$X = \frac{F \times 360}{N} \text{ (para los datos directos)}$$

$$x = \frac{\% \times 360}{100} \text{ (para los porcentajes)}$$

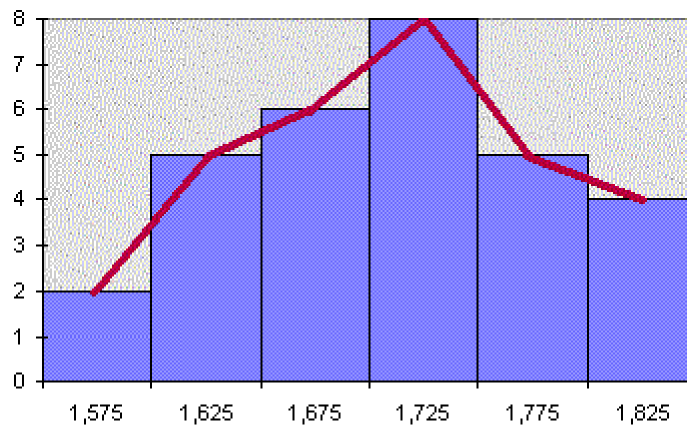


Histograma: Se trata de una serie de rectángulos que tienen sus bases sobre el eje horizontal, con centros en los puntos medios de los intervalos y anchura igual a la amplitud de los intervalos. En el eje vertical se colocan las frecuencias, proporciones o porcentajes, siempre especificándolo. Si tomamos las frecuencias, proporciones o porcentajes acumulados se denomina *histograma de frecuencias acumuladas*..



Polígono de frecuencias: En el caso de que la variable sea continua se suele preferir el polígono de frecuencias. Se construye sobre el histograma y se sitúan unos puntos en el centro de la línea superior de cada uno de los rectángulos. Si se toman las frecuencias, proporciones o porcentajes acumulados, el gráfico se denomina *polígono de frecuencias*

acumuladas y la línea que une los distintos intervalos pasa por el límite superior de cada uno de ellos, en vez de hacer por sus puntos medios. El polígono de frecuencias es una representación gráfica más sencilla que su correspondiente histograma.



4. MEDIDAS DE TENDENCIA CENTRAL

Los estadísticos de tendencia central buscan un valor que sirva par representar a los sujetos de la muestra.

4.1. Moda (Mo)

Es la categoría o valor que se repite más veces en una distribución. Es decir, aquel valor que tiene mayor frecuencia. Es la más imperfecta de las tres medidas de tendencia central.

Ejemplo: 1,1,5,5,6,<u>7,7,7</u>,8,9, ... Mo = 7

Para determinar la moda es necesario que antes se ordenen y se tabulen los datos, para conocer la frecuencia de cada valor o intervalo. Puede suceder que en algunas distribuciones se encuentre más de una moda. Así, podemos hablar de

Distribución unimodal	2,4,6,6,8,10 ....Mo = 6
Distribución multimodal	1,1,1,5,6,6,6,7,8,8 ....Mo = 1 y Mo = 6
Distribución amodal	2,4,6,8,10

Cuando los dos valores que se repiten igual número de veces son contiguos, sólo habrá una moda:

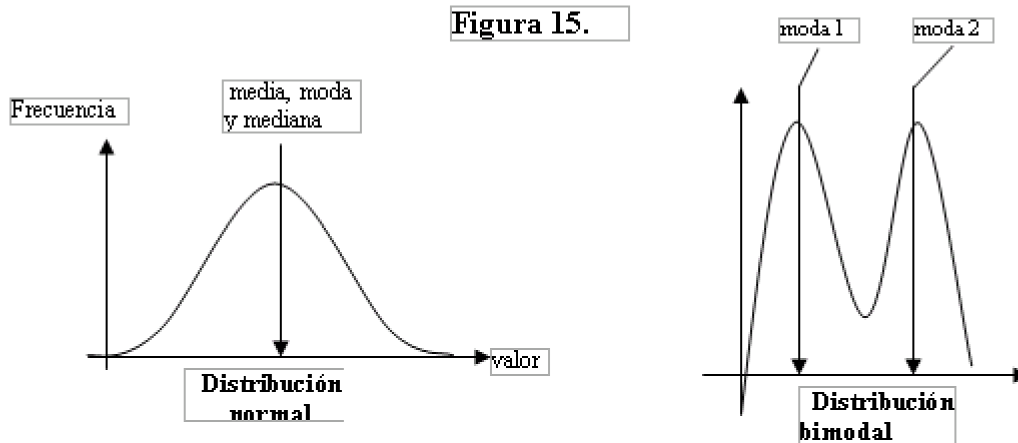
Ejemplo: 1,1,5,5,6,6,6,7,7,7 ....Mo = $\frac{6 + 7}{2}$ = 6'5

En el caso de que la variable medida se agrupe en intervalos, hay que referirse al "intervalo modal" que es el intervalo que tiene mayor frecuencia. La Mo será la marca de clase de dicho intervalo.

130 - 136	6
123 -129	5
116 - 122	**10**
109 - 115	8
102 - 108	9
95 - 101	6

Mo = 119 .... 116 + 122 / 2 = 119

Puede que en una distribución haya más de una moda, pero que no todas tengan la misma altura. La más alta sería la *moda mayor* y la/s otra/s sería la *moda menor*.



Propiedades de la moda

- Es un estadístico muy inestable por lo que será más conveniente su cálculo cuando tengamos muestras grandes.
- La moda es la única medida de tendencia central que puede calcularse cuando trabajamos en una escala nominal y otras escalas superiores.
- En el caso de que haya intervalos, el valor de la moda dependerá de los intervalos que hayamos formado.
- Cuando la moda cae en un valor extremo de la distribución, entonces no puede considerarse representativa de la tendencia central.
- La moda puede calcularse aunque una distribución tenga intervalos abiertos, siempre que no caiga en uno de esos intervalos.

4.2. Mediana (Md)

Es la puntuación central de una serie de datos ordenados; es la puntuación que deja por debajo el 50% de los datos y también por arriba. Es el dato que queda en el centro de la distribución cuando los datos están ordenados de menor a mayor. Sólo se puede utilizar a partir de escalas ordinales.

Cálculo

1º Se calcula la posición de la Mediana: Posición = $\frac{N + 1}{2}$

2º Se busca el valor que ocupa dicha posición (Md)

Ejemplo: 3,3,5,**5**,7,7,7 Posición = $\frac{N + 1}{2} = \frac{7 + 1}{2} = 4$ ....Md = 5

- En el caso en que se trate de un número par de datos, se calcula la posición y después se realiza la semisuma de las dos puntuaciones más próximas al centro.

Ejemplo: 3,3,5,**5,7**,7,7,8 Posición = $\frac{8 + 1}{2} = 4'5$ ....Md = $\frac{5 + 7}{2} = 6$

- Cuando los datos están agrupados en una distribución de frecuencias se utiliza la siguiente fórmula:

$$Md = L_i + \left(\frac{N/2 - F_{ai}}{F} \right) A$$

Donde:

L_i = límite real inferior del intervalo en el cual caiga la mediana (1ª frecuencia que sea mayor que N / 2).

N = número total de observaciones

F_{ai} = frecuencia acumulada del intervalo anterior a aquel en que cae la mediana.

F = frecuencia absoluta del intervalo en el que cae la mediana.

A = Tamaño del intervalo (amplitud) en el que cae la mediana.

Ejemplo:

INTERVALOS	F	F_a
128 - 134	15	88
121 – 127	21	73
114 - 120	**23**	**52**
107 - 113	17	29
100 - 106	12	12

1) Se calcula la posición: Posición = $\frac{N}{2} = \frac{88}{2} = 44$

2) El intervalo es el 114 – 120

L_i = 113’5

F = 23

$$Md = 113{,}5 + \left(\frac{44 - 29}{23} \right) 7 = 118{,}06$$

A = 120 – 114 + 1 = 7

F_{ai} = 29 (del intervalo anterior)

F = 23 (del intervalo en el que cae la Md)

Propiedades de la Mediana

- Es un estadístico resistente o robusto ya que no se ve afectado por todas las puntuaciones de la serie de datos. Esto también quiere decir que no es máximamente eficiente.
- Es una buena medida de tendencia central cuando tenemos puntuaciones extremas.
- Es un buen estadístico cuando la distribución es muy asimétrica porque no se ve tan afectada como la media por valores extremos.
- La Mediana puede calcularse aunque existan intervalos abiertos, siempre que no caiga en uno de esos intervalos.
- Depende de la distribución de frecuencias.
- Si tenemos varias distribuciones, cada una con su Md, la Mediana de la distribución total será mayor o igual que la Md menor, y menor o igual que la Md máxima.

Es decir: $Md_{menor} \leq Md_{total} \leq Md_{mayor}$

Ejemplo:

A = 7,7,8 ....$Md_A = 7$

B = 5,6,6,7,7 .....$Md_B = 6$ $2 \leq Md_{TOTAL} = 6 \leq 7$

C = 1,1,2,3,4 .....$Md_C = 2$

1,1,2,3,4,5,**6**,6,7,7,7,7,8 .... Posición = $\frac{13+1}{2}$ = 7 .... Md = 6

- Debido a que la Md es una posición, se deben ordenar los datos antes de llevar a cabo cualquier cálculo. Esto implica consumo de tiempo para cualquier conjunto de datos que contenga un gran número de elementos (intervalos o medidas directas). Por consiguiente, si se desea utilizar una medida de tendencia central en tales casos, es más sencillo usar la media.

4.3. Media aritmética ($\overline{X}$)

Es la más utilizada tradicionalmente en las ciencias porque es muy eficiente. Capta la variación de cada una de las puntuaciones. La media aritmética de la muestra se representa con $\bar{x}$, mientras que la media aritmética de la población se representa con la μ.

Cáculo

- *En el caso de datos no agrupados en intervalos o puntuaciones directas:*

$$\overline{X} = \frac{\Sigma x_i}{N} \text{ (cuando la frecuencia de } x_i = 1) \quad \text{ó} \quad \bar{x} = \frac{\Sigma x_i F_i}{N}$$

	X_i	F	F_a	$X_i F_i$
	35	2	34	70
	30	3	32	90
	25	5	29	125
	20	7	24	140
	15	9	17	135
	10	5	8	50
	5	3	3	15
Σ =		34		625

$$\bar{x} = \frac{\Sigma x_i F_i}{N} = \frac{625}{34} = 18{,}38$$

- *En el caso de datos agrupados en intervalos:*

$$\overline{X} = \frac{\Sigma(x_m F_i)}{N}$$

Donde:

Σ = Sumatorio

x_i = puntuaciones

F_i = frecuencia de cada puntuación

N = nº de observaciones de la muestra

X_m = marca de clase de cada intervalo

Ejemplo:

x_i	F	x_m	$X_m F_i$
15 – 25	378	20	7560
25 – 45	324	35	11340
45 - 75	108	60	6480
Σ =	810		25380

$$\overline{X} = \frac{\Sigma(x_m F_i)}{N} = \frac{25.380}{810} = = \mathbf{31'33}$$

Propiedades de la media aritmética

- Se puede decir que la media, en términos físicos, es el centro de gravedad de la distribución.
- Es un estadístico máximamente eficiente y mínimamente robusto.
- No es un buen estadístico de tendencia central cuando en la distribución hay valores extremos o cuando es muy asimétrica.
- Cada conjunto tiene una y sólo una media.
- Si multiplicamos o dividimos todas las puntuaciones de una serie de datos por una constante, la $\overline{x}$ de las nuevas puntuaciones será igual a la $\overline{x}$ de las antiguas por o dividido por la constante.

$\overline{y} = c \cdot \overline{x}$ Ej. x = 7,3,2,4 .... $\overline{x} = 4$

c = 2

y = 14,6,4,8 .... $\overline{y} = 2 \times 4 = 8$

- Si sumamos o restamos una constante a todas las puntuaciones de una distribución, la x de los nuevos datos será igual a la x original más o menos la constante.

$\overline{y} = x + c$ Ej. x = 7,3,2,4 .... $\overline{x} = 4$

c = 2

y = 9,5,4,6 .... $\overline{y} = 4 + 2 = 6$

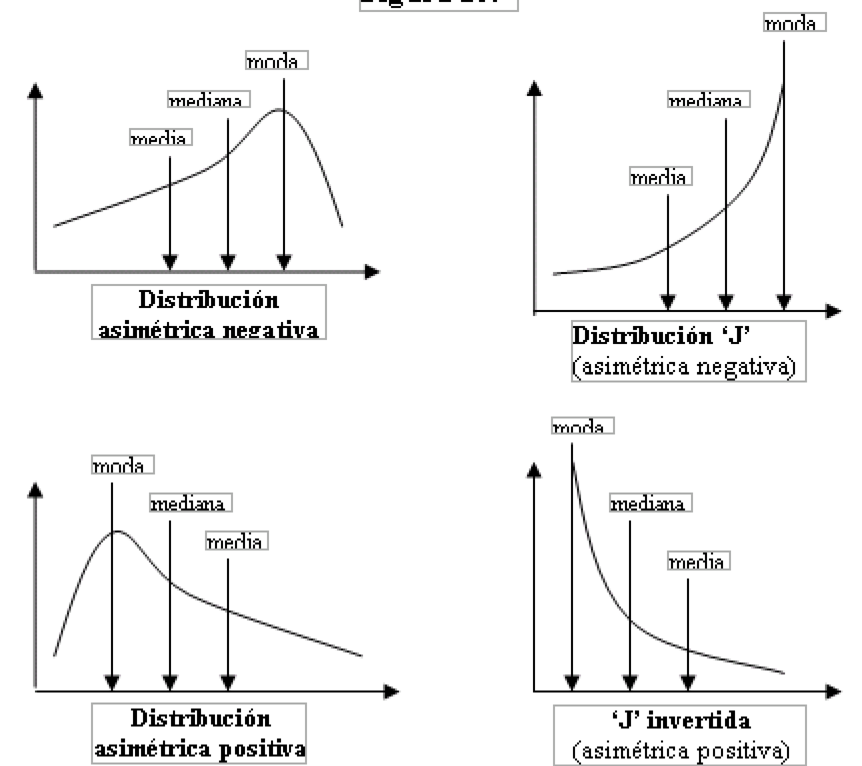
Estas dos propiedades se pueden sintetizar de la siguiente manera:

$$\overline{y} = Ax + B$$

4.4. Comparación entre la media, mediana y la moda

En las distribuciones simétricas	Sólo contienen una moda. Tienen el mismo valor para la $\overline{x}$, Md y Mo
En una distribución positivamente sesgada (hacia la derecha)	La Mo se encuentra en el punto más alto de la distribución, la Md está hacia la derecha de la Mo y la x se encuentra más a la derecha de las dos anteriores. La **Md** resulta ser la mejor medida de posición, ya que siempre está entre la moda y la media.
En una distribución negativamente sesgada (hacia la izquierda)	La Mo sigue siendo el punto más alto. La Md está hacia la izquierda de aquélla y la media más a la izquierda de la moda y la mediana. La **Md** resulta ser la mejor medida de posición, debido a que siempre está entre la moda y la media. La mediana no se ve influenciada por la frecuencia de aparición de un solo valor, como la Mo, ni se distorsiona con la presencia de valores extremos como la media.

- Distribución simétrica: Mo = Md = $\overline{x}$
- Distribución asimétrica positiva: Mo < Md < $\overline{x}$
- Distribución asimétrica negativa: Mo > Md > $\overline{x}$



5. MEDIDAS DE VARIABILIDAD O DISPERSIÓN

Los *estadísticos de variabilidad* son los que nos permiten tener una idea de la dispersión de los datos respecto a algún valor promedio. Se refiere a la extensión de los datos de una distribución.

5.1. Amplitud Total (AT)

Es la primera aproximación que podemos hacer para calcular la variabilidad de los datos. Consiste en la distancia entre el valor máximo y el mínimo. Es, además, el más inexacto.

En el caso de datos originales:
Amplitud total excluyente: AT_E = Punt. Máx. – Punt. Mín. (una de las dos puntuaciones no está incluida)
Amplitud total incluyente: AT_I = Punt. Máx. – Punt. Mín. + 1

Ejemplo: x = 2,5,7,4,2,**1**,3,10,9,11,**12**,9

$AT_E = 12 – 1 = 11$

$AT_I = 12 – 1 + 1 = 12$

En el caso de intervalos:

Amplitud total excluyente (AT_E) es igual a los datos originales

Amplitud total incluyente: AT_I = Punt. Máx. – Punt. Mín. + Amplitud del intervalo

ó

AT_I = Lím real sup. – Lim real inf.
(intervalo máx.) (intervalo mín.)

Ejemplo:

i	**F**	**x_i**
10 – 12	4	11
7 – 9	7	8
4 – 6	10	5
1 - 3	8	2

$AT_E = 11 – 2 = 9$

$AT_I = 11 – 2 + 3 = 12$ ó AT_i = 12’5 – 0’5 = 12

5.2. Los Cuantiles (medidas de posición)

Nos indican la posición relativa que ocupa una determinada puntuación de la cual sabemos el porcentaje de casos que deja por arriba y por abajo. Existen diversos tipos:

***Cuartiles* (Q)**

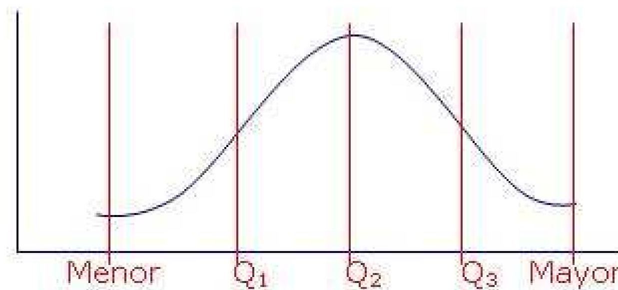
Son las puntuaciones que dividen a una distribución en cuatro partes iguales:

Q_1 = Es el que deja por debajo de sí una cuarta parte de las puntuaciones, y por encima las tres cuartas partes restantes. Es decir, deja por debajo el 25% y por encima el 75% de las observaciones.

Q_2 = Se encuentra en el punto medio, dejando por debajo el 50% de los datos y por encima el otro 50%.

Q_3 = Deja el 75 % de las observaciones por debajo, y el 25% por encima.

Q_4 = Deja el 100% de los datos por debajo. Este cuartil no se suele utilizar.



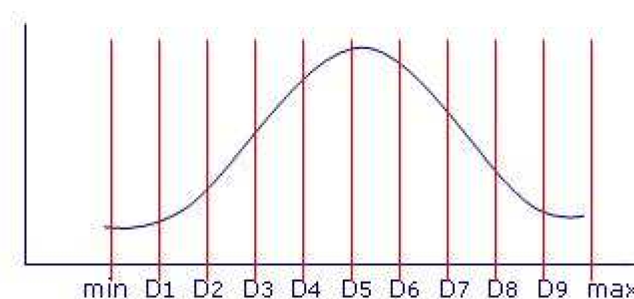
***Deciles* (D)**

Dividen el conjunto de datos en diez partes iguales. De esta manera,

D_1 = Dejaría el 10% o 1/10 parte de los datos por debajo y el 90% por encima.

D_5 = Deja el 50% de las observaciones tanto por debajo como por encima.

Y así sucesivamente.



***Percentiles* (P) *o Centiles* (C)**

Son los cuantiles que dividen el conjunto de datos en cien partes iguales, cada una de ellas con el 1% de los datos. Por ejemplo, el P_{33} dejaría por debajo el 33% de los datos.

Los cuartiles y los deciles son equivalentes a los percentiles. Por ejemplo:

30% = $P_{30} = D_3$; 75% = $P_{75} = Q_3$; 50% = $P_{50} = Q_2 = D_5$ = Md

Pasos para el cálculo de los percentiles:

1) $pn = \frac{p}{100} \cdot N$ donde p es la posición del percentil que queremos conocer. P_{60}…p=60

2) $P_p = L_i + \frac{(pn - F_{ai})}{F} A$

5.3. La Varianza (s^2) y la desviación típica (s)

La idea para su construcción surge de cuantificar las distancias, y por tanto la variabilidad, entre los valores de la variable a través de su diferencia respecto de una medida central como es la media.

$$\text{Varianza} = s^2 = \frac{\Sigma(x - \overline{x})^2}{N}$$

Debido a que las unidades de la varianza son las de la variable pero al cuadrado, se define la desviación típica o estándar como la raíz cuadrada de la varianza.

$$\text{Desviación típica } = s = \sqrt{\frac{\Sigma(x - \overline{x})^2 F}{N}} \quad \text{ó} \quad s = \sqrt{s^2}$$

Cuando trabajemos con intervalos x = la marca de clase de cada intervalo.

Cuando la varianza y la desviación típica hacen referencia a la población los parámetros que se utilizan son σ^2 para la varianza y σ para la desviación típica.

i	**F**	**x_i**	**x_i F**	**$(x – x)^2$ F**
25 – 27	3	26	78	147
22 – 24	5	23	115	80
19 – 21	10	20	200	10
16 – 18	8	17	136	32
13 - 15	4	14	56	100
10 - 12	2	11	22	128
	32		607	497

$\overline{x} = 607 / 32 = 18{,}9 \approx 19$

$s^2 = 497 / 32 = 15{,}53$

$s = \sqrt{15{,}53} = 3{,}94$

Propiedades de la varianza y la desviación típica

- 0 sería la variabilidad mínima, no podemos encontrar variabilidad negativa.
- No hay un valor máximo, por ello no es fácil de interpretar. Lo que se hace son comparaciones entre distribuciones, en estados de variabilidad.
- Como se basan en el cálculo de la media son máximamente eficientes, pero se ven afectados por todas las puntuaciones de la distribución. Su cálculo es adecuado cuando lo es la media como medida de tendencia central.
- La s viene expresada en las mismas unidades que los datos, mientras que la s^2 en unidades cuadráticas.

5.4. Coeficiente de variación (CV)

Es un indicador de la variabilidad relativa de los datos. Se trata de un porcentaje que afirma la mayor o menor variabilidad de los datos. Es decir, lo que pretende medir es el porcentaje de variabilidad respecto al total, lo cual permite comparar datos.

$$\mathrm{CV} = \frac{s}{\bar{x}} 100$$

A partir del ejemplo anterior: $\mathrm{CV} = \frac{3{,}94}{19} 100 = 20{,}73\%$

Propiedades del coeficiente de variación

- En sentido estricto sólo puede explicarse cuando trabajamos en escala de razón. A nivel práctico también en la de intervalos.
- Se trata de un valor abstracto; no tiene unidades de medida.
- No es una buena medida cuando $\bar{x} = 0$ o un valor próximo a 0.

5.5. Estadísticos de variabilidad y escalas de medida

E. Nominal	E. Ordinal	E. Intervalos	E. Razón
AT	AT	AT	AT
	Q	Q	Q
		s	s
		s^2	s^2
			CV

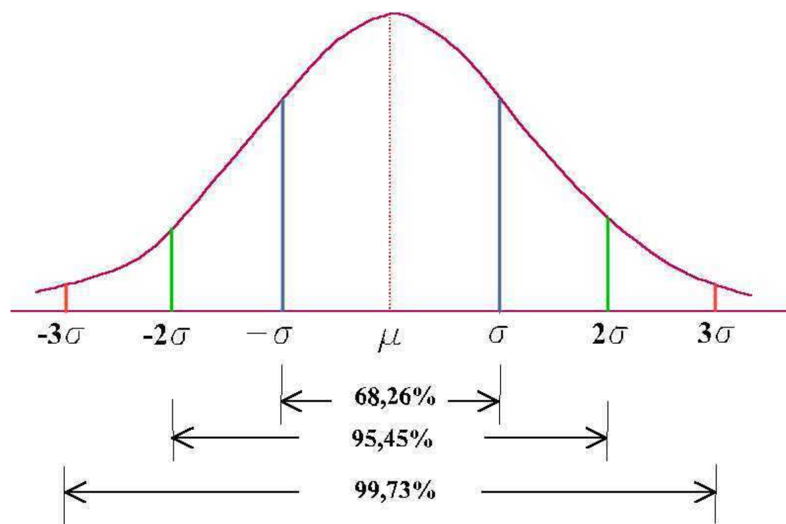
6. LA CURVA NORMAL

6.1. Concepto y características de la curva normal

Existen diferentes tipos de distribuciones, aunque la más utilizada es la **distribución normal o gaussiana**, ya que la mayoría de los eventos de la naturaleza se distribuyen de esa forma. Fue desarrollada por Gauss, de ahí que se la conozca como *curva normal* o *campana de Gauss*.

Propiedades

- Todas las distribuciones normales son simétricas por la media y, cuando tenemos un curva normal "perfecta", coincide con la mediana y la moda: $\bar{x} = \text{Md} = \text{Mo}$. La media se situaría en el centro de la curva.
- Tienen dos puntos de inflexión en $\bar{x} \pm s$: el punto donde la curva pasa de ser cóncava a ser convexa.
- Entre la $\bar{x} \pm 1s$ se encuentran el 68'26% de los datos, entre la $\bar{x} \pm 2s$ el 95'44%, y entre la $\bar{x} \pm 3s$ están el 99'74% de las puntuaciones.
- La curva tiene un solo pico, es decir es unimodal. Tiene forma de campana.



Ventajas y utilidad

- En las variables que tienen en la población una distribución gaussiana o normal se puede predecir su comportamiento en la población. Es decir, desde los valores de la muestra se pueden inferir los resultados para la población, ya que el

modelo gaussiano permite aplicar teoremas matemáticos que facilitan estas inferencias.

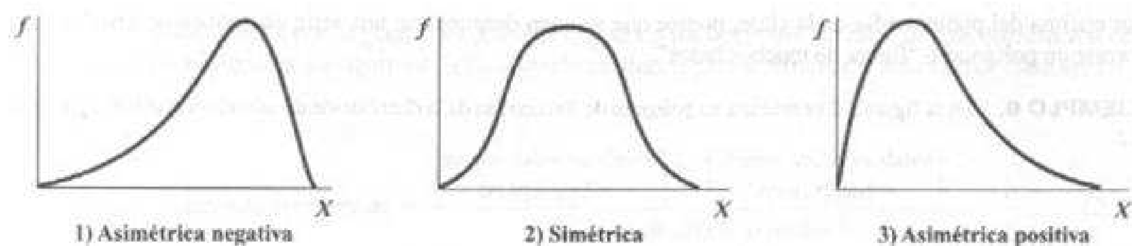
- A partir de un punto o valor en el eje de abcisas, se puede calcular la proporción de valores que quedan por debajo o por encima. Dado que los valores o puntuaciones directas se pueden convertir en puntuaciones típicas, algunos autores han elaborado tablas que determinan tales proporciones a partir de un punto de la curva normal.

6.2. Concepto de asimetría. Tipos

Se dice que una distribución es **simétrica** si al dividir en dos partes, con igual superficie, mediante la Md, una parte es imagen de la otra. Una distribución será **asimétrica** cuando presenta algún sesgo.

Tipos de asimetría

- *Positiva:* cuando presenta un sesgo positivo, es decir, hacia la derecha.
- *Negativa:* cuando el sesgo es negativo, es decir, hacia la izquierda.



Como se comentó anteriormente:

- En una distribución simétrica: Mo = Md = $\bar{x}$
- En una distribución asimétrica positiva: Mo < Md < $\bar{x}$
- En una distribución asimétrica negativa: Mo > Md > $\bar{x}$

6.3. Puntuaciones transformadas

En ocasiones, no podemos comparar varias muestras utilizando estadísticos de posición como los percentiles o los cuartiles. Los percentiles se caracterizan por ser una escala ordinal siendo poco útiles a la hora de hacer cálculos estadísticos. Para ello, es

conveniente transformar los datos en otro tipo de puntuaciones en escala de intervalo, como son las puntuaciones típicas.

***Puntuaciones directas* (X, Y, Z)**

Se obtienen directamente cuando hacemos una medición. Se simbolizan con las letras latinas mayúsculas. Este tipo de puntuaciones por sí solas aportan poca información. Para conocer el significado de una puntuación hay que ponerla en relación con el resto de puntuaciones. En esto consisten las puntuaciones diferenciales.

***Puntuaciones diferenciales o desviadas* (x, y,..)**

Se obtienen a partir de la diferencia entre las puntuaciones directas y la $\bar{x}$: $x = X - \bar{X}$. Se simbolizan con letras latinas en minúsculas.

Características:

- $\Sigma x = 0$ .... $\Sigma(x - \bar{x}) = 0$
- La $\bar{x} = 0$ .... $\Sigma \times f / N = 0$
- La $s_x = s_X$
- Las puntuaciones positivas indican valores directos mayores a la media.
- Las puntuaciones negativas indican valores directos menores a la media.
- Las unidades de la nueva escala son unidades de medida de desviación.

Las puntuaciones diferenciales aportan información de las puntuaciones con respecto a la media, pero no aportan información sobre la magnitud de esa diferencia, de manera que no son comparables entre sí las puntuaciones diferenciales entre distintas muestras.

***Puntuaciones típicas* (z)**

Las puntuaciones z o típicas comparan las diferencias con la desviación típica. Expresan el número de desviaciones típicas que un determinado valor se desvía de la media y la dirección de esa desviación. Tiene en cuenta no sólo la media sino la variabilidad del grupo. Se obtiene de la siguiente manera:

$$z = \frac{x}{S_x} = \frac{X - \overline{X}}{S_x}$$

A la transformación de las puntuaciones directas en puntuaciones típicas se le llama *tipificación de la variable*. Este tipo de puntuaciones permite llevar a cabo comparaciones entre grupos o muestras diferentes.

Características:

$\Sigma z = \frac{\Sigma x}{s}$ .... $\frac{1}{s}\Sigma x = 0$ .... $\Sigma z = 0$

La media de z es igual a 0 .... $\overline{z} = 0$

$\Sigma z^2 = N$ .... $\Sigma z^2 F = N$

$S_Z = 1$ .... $s_z = \sqrt{\frac{\Sigma(z - \overline{z})^2}{N}} = \sqrt{\frac{\Sigma z^2}{N}} = \sqrt{\frac{N}{N}} = 1$

Ejemplo:

$X_i = 90$

$\overline{X} = 70$ $\qquad\qquad z = \frac{90 - 70}{10} = \frac{20}{10} = 2$

S = 10

La puntuación directa 90 está 2 desviaciones típicas por encima de la media. Si fuera una valor negativo estaría por debajo de la media.

Para conocer el porcentaje de sujetos que se quedarían por encima o por debajo de una determinada puntuación, se utilizan unas tablas para distribuciones normales, en las que, una vez conocida la puntuación z, podemos conocer el porcentaje (fotocopia).

En el ejemplo anterior y tras consultar la tabla sabremos que si:

$z = 2$ → $P(z < 2) = 0.9772$ → El 97.72% de los sujetos quedarían por debajo de la puntuación 90. Es decir, habrían obtenido una puntuación inferior a 90.

BIBLIOGRAFIA

Amón, J. (1999): *Estadística para psicólogos I. Estadística descriptiva*. Madrid, España: Pirámide.

Botella, J., León, O. G., San Martín, R. y Barriopedro, M. I. (2001). *Análisis de datos en psicología I. Teoría y ejercicios*. Madrid, España: Pirámide.

Pérez Juste, R., García Llamas, J. L., Gil Pascual, J. A. y Galán González, A. (2009). *Estadística aplicada a la educación*. Madrid, España: UNED/Pearson-Prentice Hall.

Selva, J., Cervera, T., Dasí, C., Ruiz, J. C. y Meliá, J. L. (1991). *Problemas de psicoestadística descriptiva*. Valencia, España: Cristóbal Serrano.